**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Team Members:   1. Kunal Mahadik   Email id : kunalmahadik0811@gmail.com  Contribution:   1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Decision Trees and Random Forest. 6. Aashruti Agarwal   Email id : aaashruti@gmail.com  Contribution:   1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of KNN algorithm. 6. Raneev K   Email id : raneevk36@gmail.com  Contribution:   1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Logistic Regression. |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/Raneevk/Credit-Card-Default-Prediction-capstone-project |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Lending is one of the key business areas in the banking industry, credit cards as of late have seen huge success over the course of years. In pursuit to increase their market share, banks often issue credit cards to ineligible customers without adequate background checks. Also, many customers used their credit card beyond their repayment capabilities leading to high debt accumulation. Identifying the risky and non-risky customers is the biggest challenge for banks. So, the problem we are trying to analyze is how to identify the risky and non-risky customers, helping the bank to decide if a customer has the potential to repay the used credit of the bank.  First of all we load the dataset and performed data cleaning and null value treatment. Thankfully there were no null values and duplicate values are present in the dataset. We did outlier treatment by using IQR caping method and did some EDA and data visualization. We plot different graphs of univariate analysis and bivariate analysis and make different inference from our dataset.  At the end part of EDA we plot the correlation heatmap and find the correlation of each independent variable with our target variable. There was class imbalance in our target variable ie. More than 50% difference between both the classes .To treat class imbalance we used SMOTE technique while doing modeling.  In feature engineering part we used ExtraTree classifier to check the feature importance of each variable i.e which feature is more important compared to our model and which is of less importance. Then we used ANOVA to select the best feature which we will be using further in our model.  In modeling part we tried four different classification models which are logistic regression, Decision tree classifier, Random Forest classifier and KNN.  CONCLUSIONS:   1. Logistic Regression had an imbalance in the recall score of about 83% for class 0 and 56% for class 1. 2. Performance on Decision Tree and Random Forest is comparatively better. Decision Trees and Random Forest have recall scores of 75%(class 0) , 49%(class 1) and 65%(class 0), 66%(class 1) respectively. 3. KNN classifier could be a good model but it needs further hypertuning. |